Exploring the Synergy Between Vision-Language Pretraining and ChatGPT for Artwork Captioning: A Preliminary Study

Giovanna Castellano[0000−0002−6489−8628], Nicola Fanelli, Raffaele Scaringi[0000−0001−7512−7661], and Gennaro Vessio[0000−0002−0883−2691]

Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
{giovanna.castellano,raffaele.scaringi,gennaro.vessio}@uniba.it
n.fanelli10@studenti.uniba.it

Abstract. While AI techniques have enabled automated analysis and interpretation of visual content, generating meaningful captions for artworks presents unique challenges. These include understanding artistic intent, historical context, and complex visual elements. Despite recent developments in multi-modal techniques, there are still gaps in generating complete and accurate captions. This paper contributes by introducing a new dataset for artwork captioning generated using prompt engineering techniques and ChatGPT. We refined the captions with CLIPScore to filter out noise; then, we fine-tuned GIT-Base, resulting in visually accurate captions that surpass the ground truth. Enrichment of descriptions with predicted metadata improves their informativeness. Artwork captioning has implications for art appreciation, inclusivity, education, and cultural exchange, particularly for people with visual impairments or limited knowledge of art.

Keywords: ChatGPT · Computer vision · Cultural heritage · Deep learning · Digital humanities · Image captioning.

1 Introduction

Artwork captioning refers to generating concise and informative text descriptions that capture the essence of an artwork, its visual elements, and underlying concepts [7]. This emerging field has significant potential not only to enrich art appreciation but also to promote inclusivity, education, and cultural exchange, particularly for people with visual impairments or limited artistic knowledge.

However, generating rich and semantically meaningful captions for artworks poses unique challenges [6]. Artistic expression often transcends literal representation, incorporating abstract concepts, emotions, and symbolic meanings. Capturing and effectively communicating these elements in textual descriptions requires a deep understanding of artistic intent, cultural references, and historical context. Moreover, the complexity of visual elements within artworks, including colors, textures, and spatial relationships, adds a layer of difficulty in generating comprehensive captions.

Multi-modal techniques for artwork captioning have gained attention in recent years, with studies exploring different approaches. For example, *neural style transfer* has been employed to transform images into paintings, creating a large-scale dataset with image-text pairs [20]. Iconographic captions and visual question answering in cultural heritage have also been explored [3, 7]. However, there are still gaps in developing more comprehensive and accurate captioning techniques, addressing the limitations of dataset design, and improving the model's knowledge and understanding of artworks.

In this paper, we aim to contribute to this research by introducing a new dataset designed explicitly for automatically generating artwork captions. Using prompt-engineering techniques, we exploited ChatGPT [21] to generate visual descriptions of artworks based on title and artist information. Although these descriptions focus on the content of the artwork and contain conceptually rich elements, there is a significant presence of noise caused by ChatGPT hallucinations, particularly for lesser-known artworks. We found that CLIPScore [15] is an effective indicator of caption noise and used it to filter out poor examples and assign weights to the remaining descriptions. We then fine-tuned a vision-language pre-trained (VLP) model, the Generative Image-to-text Transformer (GIT) in its base version [31], resulting in a new framework for creating artwork captions that generate visually accurate captions that are superior to the ground truth. In addition, we enriched the visual descriptions with predicted metadata using a multi-task classification model based on the Vision Transformer (ViT) architecture [11], improving their informativeness.

The rest of the paper is organized as follows. Section 2 reviews related literature. Section 3 presents the data used in this study. Section 4 describes the proposed methodology. Section 5 presents our experimental evaluation. Section 6 concludes the paper and discusses future directions for our research.

2 Related Work

Since their introduction in neural machine translation, Transformers [30] have found several applications in the domain of image captioning [9, 12, 14]. In particular, using Transformers has facilitated the emergence of vision-language pre-training as a powerful approach to cross-modal learning by exploiting large-scale models and datasets. VLP models are commonly pre-trained on extensive collections of unlabeled or weakly labeled multimodal data, using pretraining objectives to develop a holistic understanding of vision and language. Subsequently, these models can be fine-tuned on various downstream tasks.

However, while much work in automatic image captioning has been done in the general domain of natural images, very few studies have tackled this task in the more challenging fine arts domain, arguably one of the most problematic domains in which to perform this task, both because of its complexity and the absence of rich task-specific datasets [26]. Initially, this research focused primarily on image-text and text-image retrieval to exploit the synergy between textual and visual content to improve the effectiveness and accuracy of the search. In a

seminal paper, Garcia et al. [13] presented *SemArt*, the first dataset of fine art images paired with corresponding artistic commentaries. They conducted several experiments using this dataset, paving the way for further exploration in the field. Another significant contribution came from Stefanini et al. [27], who introduced *Artpedia*, a dataset containing paired fine art images and annotated texts. These annotations categorize the text into "contextual" and "visual" sentences.

Over time, the application of multi-modal techniques has broadened, attracting the interest of researchers working on more complex tasks, such as artwork captioning. In [2], a description generation system based on *SemArt* was proposed, which uses an encoder-decoder model (ResNet-LSTM) to generate multi-topic artwork descriptions covering content, form, and context, using placeholders instead of named entities. A parallel process performs metadata classification and object detection on the artwork image, generating prompts for DrQA [8] and using retrieved documents to fill placeholders in the generated description. Other works [19, 25] explored data-driven approaches for generating captions for ancient artworks. In another study, Lu et al. [20] employed *neural style transfer* to transform images from the MS COCO dataset into paintings, creating a large-scale image caption dataset with original MS COCO captions. Cetinic [7] explored iconographic captions using the Iconclass AI Test Set dataset, developing a VLP model to recognize iconographic elements from images of artworks. However, this dataset was not explicitly designed for captions, and ground-truth captions were generated through preprocessing steps applied to image labels. An alternative study conducted by Ruta et al. [23] presented a new dataset called *StyleBabel*, encompassing artworks from various genres. This study focused on artwork tagging and captioning. More recently, Ishikawa and Sugiura [16] approached artwork captioning from a different perspective, emphasizing the affective dimension of image captions. Recent work studied visual question answering in the cultural heritage domain, developing models using the VISCOUNTH dataset [3] or employing specific prompts with GPT-3, demonstrating model knowledge of specific and famous artworks [4].

In our research, we curated a dataset designed explicitly for the automatic generation of artwork captions. Using ChatGPT, we generated descriptions based on artists and titles of artworks from our $\mathcal{ArtGraph}$ Knowledge Graph [5]. Although the descriptions focused on the content of the artwork and included rich concepts, noise was present due to ChatGPT hallucinations, especially for lesser-known artworks. We used CLIPScore [15] to filter out poor examples and fine-tuned GIT-Base [31], a VLP model, resulting in a new captioning framework that generates visually accurate captions that overcome the ground truth. In addition, we enriched descriptions with predicted metadata using a ViT-based model [11], enabling the integration of other textual information.

3 Materials

Building a comprehensive dataset of richly annotated artwork captions poses significant challenges, requiring human effort and expertise. To overcome these dif-

ficulties, we used an innovative approach based on an artificially created ground truth derived from ChatGPT, the widely adopted chatbot. This approach not only streamlines the data collection process but also provides a unique opportunity to explore the intersection of ChatGPT and art curation.

As a starting point, we used *ArtGraph*, our recently released Knowledge Graph on art, built by scraping WikiArt and DBpedia [5]. It collects 116,475 artworks spanning 18 genres and 32 styles, and many other metadata that characterize them. However, despite incorporating semantic concepts through metadata, *ArtGraph* lacks textual descriptions necessary to train an artwork captioning model. As said, we created a synthetically generated ground truth using the popular ChatGPT to fill this gap. As shown in [4], GPT-3, which was trained on a large corpus of textual data related to several domains, including art, can produce good descriptions of artworks by exploiting the information it used during the training process. However, these capabilities do not prevent the model from generating erroneous or partially erroneous descriptions.

Similarly, we asked ChatGPT to generate text descriptions for each artwork in *ArtGraph* with the following prompt followed by a list of artworks with their titles and authors:

> *Write visual descriptions for the following artworks.*
> *RULES:*
> - *Descriptions must be between 20 and 40 tokens in length.*
> - *The content of each description should only refer to the subjects, their attributes, and the scenes depicted.*
> - *Avoid repeating the author's name or the painting's title within the descriptions.*
> - *Begin each description with the phrase 'The artwork depicts'.*
> - *List the descriptions using numbers and maintain the order of the provided artworks.*
> - *Descriptions must not include false information.*

The prompt rules were refined manually after experimenting with various configurations. It is worth noting that ChatGPT rarely complied with the token limits and the rule prohibiting providing false information. In any case, the generated descriptions include information about the subjects, their attributes, and the scene depicted, and occasionally include iconographic, formal, or emotional elements to enhance the overall appeal of the captions.

To evaluate the quality of an image-caption pair, we leveraged CLIP [22], specifically its associated CLIPScore [15]. CLIP is a deep learning model that maps images and texts into a shared embedding space. It was trained using contrastive loss over 400M image-text pairs from the Internet. For an image with visual CLIP embedding $\mathbf{v}$ and the corresponding generated caption with textual CLIP embedding $\mathbf{c}$, we computed the CLIPScore as:

$$CLIPScore(\mathbf{c}, \mathbf{v}) = \max(\cos(\mathbf{c}, \mathbf{v}), 0)$$

The scores are within the range $[0, 1]$, where higher scores indicate a higher level of semantic matching between the image and the caption. In practice, scores

typically fall within the $[0, 0.4]$ range. Hessel et al. [15] showed that CLIPScore highly correlates with human judgment on image captioning tasks. Unlike traditional metrics, it does not require reference captions, so we could use it to evaluate the quality of our ground truth examples and to automatically filter out bad image-caption pairs from our dataset, a technique that has already been used to create open image-text datasets such as LAION-400M [24], where the authors heuristically chose a value of 0.3 as a threshold for filtering out bad Internet-collected examples.

In line with the findings in [29], we employed the NLP augmentation technique known as *back-translation* to generate two additional captions for each artwork. This technique involves translating the original English caption into another language and then translating it back into English, resulting in slightly different captions. For our back-translation, we utilized the OPUS-MT translation models [28] for French and German. As a result, our final dataset comprises $116,475$ $\mathcal{ArtGraph}$ images, each associated with three English captions.

4 Methods

Our framework comprises two models: a caption generator and a metadata classifier. Both models take as input the image of an artwork without any additional information, allowing our framework to be applied to any artwork whose only information is its visual appearance. When the digitized image of an artwork is fed into our framework, it is processed by the two models in parallel. Both models use ViT-B [11] as the image encoder. The caption generator is responsible for generating a visual description of the artwork and is trained using the captions generated synthetically by ChatGPT. Simultaneously, the metadata classifier predicts the artist, genre, style, tags, and media associated with the artwork and is trained using the $\mathcal{ArtGraph}$ connections of the artwork as supervised labels. The outputs of the metadata classifier are used to populate a predefined template, which is then combined with the visual caption generated by the first model. This results in a description of the artwork that highlights both information about the artwork and a visual description. Figure 1 shows the general outline of the proposed method; the functioning of the two core models is described below.

4.1 Caption Generation

For caption generation, our method involves fine-tuning GIT-Base [31]. The model uses an encoder-decoder architecture. The encoder is ViT-B/16, initialized with CLIP weights, while the decoder is a standard Transformer decoder. The entire encoder-decoder model is pre-trained on 10M image-text examples from MS COCO, SBU, Conceptual Captions (CC3M), and Visual Genome.

In this approach, the image information of an example is embedded in the caption input tokens through linear projections of the patch embeddings generated by the ViT encoder. At each time step, the decoder generates probability

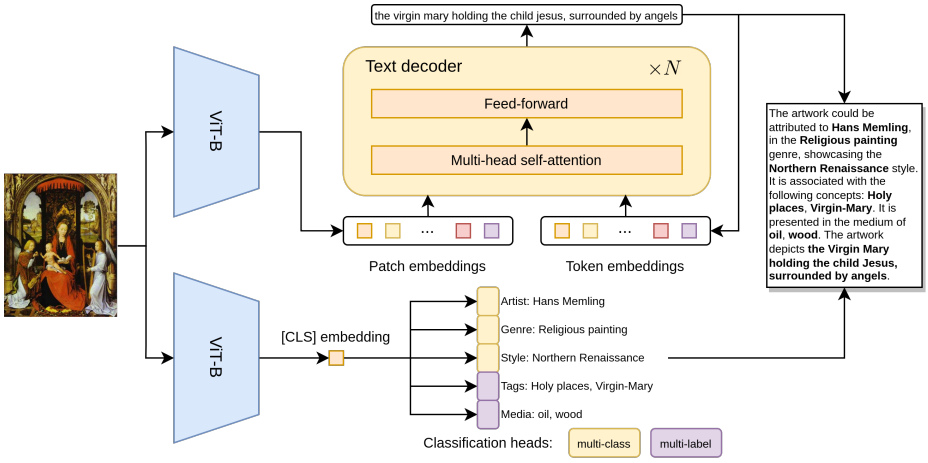


Fig. 1: Our framework includes a caption generator and a metadata classifier using ViT-B as an image encoder. The caption generator produces a visual description, trained on synthetic captions, while the metadata classifier predicts artist, genre, style, tags, and media based on $\mathcal{ArtGraph}$. The combined results provide a complete description of the artwork.

distributions over the vocabulary associated with the BERT-Base [10] uncased tokenizer, considering the context of the image. These distributions are then used to compute the language modeling loss, which is employed to train the model. Specifically, for each example corresponding to the triple (I, T, c), where I is the image, $T = t_0, t_1, t_2, \ldots, t_{N+1}$ is the sequence of caption tokens, being t_0 the `[BOS]` token and t_{N+1} the `[EOS]` token, and c is the CLIPScore computed as the cosine similarity between the CLIP embeddings of I and T, we apply the weighted variant of the loss as follows:

$$\ell = w(c)\frac{1}{N+1}\sum_{i=1}^{N+1} CE(t_i, p(t_i|I, t_0, t_1, ..., t_{i-1}))$$

where CE is the cross-entropy loss and $w(c)$ is a linear function of c, which grows proportionally as c increases, used to weight the importance of an example in the computation of the loss, based on its quality, estimated by c. This approach was designed to produce superior performance compared to the unweighted loss, as shown by the CLIPScore results in Table 2.

In all experimental settings, we excluded the final period and the initial sub-string "*The artwork depicts*" from each caption, elements that ChatGPT included during ground truth caption generation. Additionally, we converted the text inputs to lowercase to align with the same (BERT-Base uncased) tokenizer employed by GIT. Furthermore, we imposed a maximum text length of 40 tokens. To restore case information, the output of the caption generator, which is in lowercase, is modeled using *truecasing* [18].

4.2 Metadata Classification

In parallel, we fine-tuned a ViT-B/16 model, pre-trained on ImageNet-21K, to perform multitask classification on artwork images. This involved using the connections in $\mathcal{ArtGraph}$ as our ground truth. Specifically, we performed multi-class classifications for artist, genre, and style for each artwork while employing multi-label classifications for tags and media. This information is intriguing to incorporate into a visual description of an artwork, as it references the artwork's context, form, content, and style. Therefore, we injected this information into the description using a predefined textual template in which it was embedded.

To fine-tune a single image encoder, we employed a multi-classification setup by adding a linear projection of the embedding corresponding to the `[CLS]` patch as the classification head for each task. Artists with fewer than 100 associated artworks were assigned the class *other*, while media and tags with fewer than 100 associated artworks were ignored. To counteract the problem of class imbalance in multi-class classifications, each class was associated with a weight inversely proportional to its frequency in the loss calculation. Each of the five classification tasks, whether multi-class or multi-label, was associated with its own cross-entropy loss $CE_i, i \in \{1, ..., 5\}$ during training. In traditional multi-task learning, these losses are typically aggregated by summing them with empirical weights, often determined by trial and error, which can be costly and time-consuming. To overcome this problem, we opted for a more efficient approach by allowing the model to learn task weights, using the uncertainty-based approach described in [17]. By taking advantage of this method, we avoided manually adjusting the weights, making the training process more streamlined and efficient. Specifically, our model was trained using the combined loss:

$$\ell = \sum_{i=1}^{5} \frac{CE_i}{\sigma_i^2} + \sum_{i=1}^{5} \log \sigma_i$$

where σ_i^2 values represent the task variances, which are used as weights to adjust the contribution of each task to the overall loss. The model learns these weights through backpropagation (particularly, we allowed the model to learn $\log(\sigma_i^2)$ for numerical stability).

It is essential to mention that for examples without associated tags or media in the dataset, the tags and media losses were ignored. This means the model was not penalized for predicting tags or media for instances lacking these annotations, either due to missing annotations or the removal of infrequent tags or media.

5 Experiments

5.1 Experimental Setting

For all experiments, we divided the entire dataset into training, validation, and test set using a 70/15/15 stratified split on the *genre* attribute to distribute the data variability equally among the three splits. All images were treated at a

Table 1: Training configurations for caption generators (c corresponds to the instance's CLIPScore).

Model	Threshold	Train images	Learning rate	Instance weight	Encoder
GIT-Base	0.15	80,127	4.5×10^{-7}	$8c - \frac{1}{5}$	Not frozen
GIT-Base-nw	0.15	80,127	9.0×10^{-7}	No	Not frozen
GIT-Base-fr	0.15	80,127	4.5×10^{-7}	$8c - \frac{1}{5}$	Frozen
GIT-Base-gs	0.25	47,924	4.5×10^{-7}	$\frac{20}{3}c - \frac{2}{3}$	Not frozen

resolution of 224×224. We excluded examples with a CLIPScore less than 0.15 from the validation and test sets.

For caption generation, we fine-tuned several versions of GIT-Base, using a batch size of 64 (simulated using gradient accumulation). The learning rate was warmed up for the first 500 steps and followed a cosine decay to 0 for five epochs. The optimizer was AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Slight image augmentations were applied to the images (large-scale random crops, random horizontal flips and rotations, color jittering), with one epoch corresponding to three passes over the training samples, considering one of the three associated captions and an image variation as an example. All GIT-Base trainings were stopped after three epochs due to improvements of less than 1% in BLEU-1 on the validation set to save computational time.

Table 1 shows the training configurations for the tested caption generators. We conducted an ablation study to investigate the impact of different choices on the results. In addition to the basic version of GIT-Base, we trained three other variants: one without using instance weights (nw), one with frozen image encoder and word embeddings (fr), and finally, one with an increased CLIPScore threshold to exclude bad examples from the training set (gs). We also used the pre-trained version of GIT-Base without further fine-tuning on our dataset (nft) to establish an image captioning baseline for our work.

For metadata classification, we fine-tuned ViT-B/16 with five classification heads. We used a batch size of 32 (simulated using gradient accumulation), the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The encoder's weights were frozen for the first five epochs, during which only the classification heads were trained with a learning rate of 10^{-3}. After that, we unfroze the Transformer and continued training with a lower learning rate of 5×10^{-5}. We applied regularization techniques to avoid overfitting, including gradient clipping to a maximum norm of 1 and dropout just before the classification heads, with a dropout probability of 0.3. We chose to keep the model with the highest average macro F1 score across the tasks on the validation set.

The models were trained on an NVIDIA Tesla P100 GPU. Our Python implementation used PyTorch and Hugging Face for fine-tuning the pre-trained models. For evaluation, we computed BLEU-N, SPICE, METEOR, ROUGE-L, CIDEr and CLIPScore for captioning, and accuracy and macro-averaged F1 score for metadata classification on the test set predictions.

Table 2: Captioning results using greedy decoding (S: SPICE; B@N: BLEU-N; M: METEOR; RL: ROUGE-L; Cr: CIDEr; C: CLIPScore).

Model	S	B@1	B@2	B@3	B@4	M	RL	Cr	C
Ground truth	-	-	-	-	-	-	-	-	25.8
GIT-Base-nft	4.9	9.0	4.2	2.0	1.0	4.9	16.2	7.3	26.1
GIT-Base	10.0	35.3	20.7	12.4	7.6	12.0	**30.5**	31.8	26.9
GIT-Base-nw	**10.1**	**35.7**	**20.9**	**12.6**	**7.8**	**12.1**	**30.5**	**32.5**	26.6
GIT-Base-fr	9.0	35.1	20.1	11.7	7.0	11.5	29.7	27.1	26.1
GIT-Base-gs	9.9	33.9	19.8	11.7	7.2	11.6	30.0	30.1	**27.9**

Table 3: Classification results (Acc: Accuracy; F1: macro-averaged F1 score).

Model	Artist		Genre		Style		Tags	Media
	Acc	F1	Acc	F1	Acc	F1	F1	F1
ViT-B (multitask)	69.93%	58.63%	72.78%	65.94%	59.98%	57.41%	39.61%	53.55%

5.2 Results

Table 2 shows the results obtained with our GIT-Base models on the entire test set. We compare the results of all versions of GIT-Base we tested to perform an ablation study of our artwork's visual captioning method. The table shows that the best results in traditional captioning metrics are obtained from GIT-Base without weighting the instances. The worst results are obtained by freezing the ViT-B encoder and word embeddings, suggesting that it is better to let the model modify the image and word representations according to our loss. We also show that in terms of image-text matching as measured by CLIPScore, each of our models outperforms the ground truth, with the best attempt obtained by CLIPScore weighting and with a higher threshold for the selection of training samples (GIT-Base-gs). When assessing the average CLIPScore on the test set, we discovered that our top configuration (GIT-Base-gs) achieved a slightly lower value than 0.28, which is still regarded as promising. It is worth noting that Schuhmann et al. [24] established a threshold of 0.3 as a significant benchmark for image-text alignment when creating their dataset. As expected, running GIT-Base-nft, i.e., the image captioner as is, results in poor performance in relation to both traditional metrics on our dataset and CLIPScore.

Regarding metadata classification (Table 3), our results align with the genre and style classification outcomes reported in [5], albeit employing a different approach for multi-task learning. Qualitative analysis revealed that artist classification is remarkably accurate for well-known artists. Additionally, we achieved favorable F1 scores for tags and media, considering that the model can identify tags not initially added to artworks by WikiArt annotators.

A qualitative evaluation (Fig. 2) demonstrates that our GIT-Base-gs can successfully identify well-known places, personalities, and different objects and activities in a wide range of artistic styles. However, the model still experiences hallucinations resulting from the noisy ground truth. For example, the model

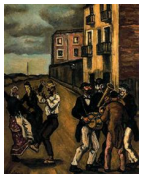
 **Rue Tronchet la Madeleine** Antoine Blanchard	 **Harlequin** Pablo Picasso	 **The Lamentation** Gerard David	 **Masks** Jose Gutierrez Solana
Rue Tronchet la Madeleine, a busy street in Paris with its signature buildings. (0.28)	*A Harlequin wearing a hat, with his arms crossed.* (0.28)	*The Lamentation, with Mary holding the body of Jesus while other figures mourn around them.* (0.30)	*Three grotesque masks with exaggerated features and vivid colors, on a neutral background.* (0.21)
A street scene in front of the Capitol building in Washington D.C. (0.26)	*A woman with a guitar in her hair and a hat.* (0.24)	*Death by painting artist depicting the death of builder.* (0.25)	*A street scene in the city of Rome, Italy. The building in front is the house of the artist.* (0.24)
The artwork could be attributed to ***Antoine Blanchard****, in the* ***Cityscape*** *genre, showcasing the* ***Impressionism*** *style. It is associated with the following concepts:* ***houses-and-buildings****,* ***roads-and-vehicles****,* ***streets-and-squares****. It is presented in the medium of* ***canvas****,* ***oil****. The artwork depicts* ***a street scene in Paris, with people and buildings in the background****.* (0.26)	*The artwork could be attributed to* ***Pablo Picasso****, in the* ***Portrait*** *genre, showcasing the* ***Cubism*** *style. It is associated with the following concepts:* ***female-portraits****. It is presented in the medium of* ***canvas****,* ***oil****. The artwork depicts* ***a self-portrait of a man with a serious expression, wearing a hat and holding a guitar in his hand****.* (0.28)	*The artwork could be attributed to* ***Albrecht Durer****, in the* ***Religious painting*** *genre, showcasing the* ***Northern Renaissance*** *style. It is associated with the following concepts:* ***Christianity****,* ***Jesus-Christ****,* ***Mythology****,* ***saints-and-apostles****. It is presented in the medium of* ***oil****. The artwork depicts* ***the Virgin Mary and the child Jesus, surrounded by a group of people and trees, with a landscape in the background****.* (0.27)	*The artwork could be attributed to* ***Pablo Picasso****, in the* ***Genre painting*** *genre, showcasing the* ***Expressionism*** *style. It is presented in the medium of* ***canvas****,* ***oil****. The artwork depicts* ***a group of musicians playing music in a street with buildings in the background****.* (0.30)

Fig. 2: Captioning examples. We present captions from our ChatGPT-generated ground truth (first row), GIT-Base without fine-tuning using our dataset (second row), and our GIT-Base-gs, which includes predicted metadata aggregation. Alongside each caption, we provide the corresponding image's CLIPScore (for the last row, computed solely based on the visual caption).

may identify objects that are not present. Additionally, the model can generate completely inaccurate captions when the content of the painting is too chaotic.

6 Conclusion & Future Work

This paper introduced a new framework for artwork captioning. Our approach involves training a VLP model on the visual descriptions of artworks generated by ChatGPT. Moreover, it improves these descriptions by incorporating predicted metadata from $\mathcal{A}rt\mathcal{G}raph$, which provides valuable information about the artwork's style, form, and context. Our research shows that accurate captions can be generated by using instance filtering, loss weighting with CLIPScore, and leveraging the prior knowledge of a VLP model such as GIT-Base.

To further improve artwork captions, we can focus on several key areas. First, it is critical to improve the quality of the dataset used. Second, we can improve the accuracy of captions by using VLP models with more parameters, and pretrained on a more extensive collection of images and texts. Another possibility for improvement is the incorporation of contextual information into captions. This can be achieved by integrating external knowledge through template-based or context-based approaches. Template-based approaches use predefined structures to include contextual information, as demonstrated in a previous paper [2], while context-based approaches directly integrate external knowledge during the caption generation process. Finally, adding emotional information to captions can significantly increase reader engagement [1].

Acknowledgment The research of Raffaele Scaringi is funded by a Ph.D. fellowship within the framework of the Italian "D.M. n. 352, April 9, 2022" - under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 3.3 - Ph.D. Project "Automatic analysis of artistic heritage via Artificial Intelligence", co-supported by "Exprivia S.p.A." (CUP H91I22000410007).

References

1. Aslan, S., Castellano, G., Digeno, V., Migailo, G., Scaringi, R., Vessio, G.: Recognizing the emotions evoked by artworks through visual features and knowledge graph-embeddings. In: Image Analysis and Processing. ICIAP 2022 Workshops. pp. 129–140. Springer (2022)
2. Bai, Z., Nakashima, Y., Garcia, N.: Explain me the painting: Multi-topic knowledgeable art description generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5422–5432 (2021)
3. Becattini, F., Bongini, P., Bulla, L., Bimbo, A.D., Marinucci, L., Mongiovì, M., Presutti, V.: VISCOUNTH: A Large-Scale Multilingual Visual Question Answering Dataset for Cultural Heritage. ACM Trans. Multimedia Comput. Commun. Appl. (apr 2023), just Accepted
4. Bongini, P., Becattini, F., Del Bimbo, A.: Is GPT-3 All You Need for Visual Question Answering in Cultural Heritage? In: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I. pp. 268–281. Springer (2023)
5. Castellano, G., Digeno, V., Sansaro, G., Vessio, G.: Leveraging knowledge graphs and deep learning for automatic art analysis. Knowledge-Based Systems **248**, 108859 (2022)
6. Castellano, G., Vessio, G.: Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. Neural Computing and Applications **33**(19), 12263–12282 (2021)
7. Cetinic, E.: Towards generating and evaluating iconographic image captions of artworks. Journal of Imaging **7**(8), 123 (2021)
8. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading Wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051 (2017)
9. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Fang, Z., Wang, J., Hu, X., Liang, L., Gan, Z., Wang, L., Yang, Y., Liu, Z.: Injecting semantic concepts into end-to-end image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18009–18019 (2022)
13. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)

14. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: Transforming objects into words. Advances in neural information processing systems **32** (2019)
15. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
16. Ishikawa, S., Sugiura, K.: Affective Image Captioning for Visual Artworks Using Emotion-Based Cross-Attention Mechanisms. IEEE Access **11**, 24527–24534 (2023)
17. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
18. Lita, L.V., Ittycheriah, A., Roukos, S., Kambhatla, N.: tRuEcasIng. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. pp. 152–159. Association for Computational Linguistics, Sapporo, Japan (Jul 2003)
19. Liu, F., Zhang, M., Zheng, B., Cui, S., Ma, W., Liu, Z.: Feature fusion via multi-target learning for ancient artwork captioning. Information Fusion **97**, 101811 (2023)
20. Lu, Y., Guo, C., Dai, X., Wang, F.Y.: Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training. Neurocomputing **490**, 163–180 (2022)
21. OpenAI: ChatGPT. https://openai.com (2023), version 3.5
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
23. Ruta, D., et al.: StyleBabel: Artistic Style Tagging and Captioning. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII. pp. 219–236. Springer (2022)
24. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
25. Sheng, S., Moens, M.F.: Generating Captions for Images of Ancient Artworks. In: Proceedings of the 27th ACM Int. Conference on Multimedia. p. 2478–2486. MM '19, Association for Computing Machinery, New York, NY, USA (2019)
26. Sirisha, U., Chandana, B.S.: Semantic interdisciplinary evaluation of image captioning models. Cogent Engineering **9**(1), 2104333 (2022)
27. Stefanini, M., Cornia, M., Baraldi, L., Corsini, M., Cucchiara, R.: Artpedia: A New Visual-Semantic Dataset with Visual and Contextual Sentences in the Artistic Domain. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) Image Analysis and Processing – ICIAP 2019. pp. 729–740. Springer International Publishing, Cham (2019)
28. Tiedemann, J., Thottingal, S.: OPUS-MT — Building open translation services for the World. In: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT). Lisbon, Portugal (2020)
29. Turkerud, I.R., Mengshoel, O.J.: Image Captioning using Deep Learning: Text Augmentation by Paraphrasing via Backtranslation. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 01–10 (2021)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: GIT: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022)